An analysis of the progression of Albania's residential housing market using microdata

THE WEB SCRAPPING METHODOLOGY

Orion Garo
Departamenti i Kërkimeve
Banka e Shqipërisë

Presentation outline

1. Introduction
2. Research question
3. Research limitations
4. Methods
5. Data details
6. Results
7. Final considerations
8. Future research

Introduction

This research study inspects the levels and dynamics of the domestic housing market main components – house prices and rentals – using microdata from the Internet.

- Objective:
  Making use of online web listings on houses for sale/rent, to identify the average levels of per-sq.m. sale/rental prices and analyze their progressions during the considered timeframe for 3 geographical divisions: capital, coastal areas, and inland areas of Albania.
- Research focus:
  Big Data - typically referring to the immense volume of Internet-residing raw microdata, notably UGC (user-generated content), with which this study is conducted.
- Underlying motivation:
  Difficult to track/estimate and to structurally analyze trends of Albania's housing market; no monthly/quarterly provision of official

Research question

What are the observed per-square-meter price averages in Albania's:

[1] residential housing market?

[2] rental market?

How do these price levels compare between the country's capital, its coastal, and its inland areas?

What structural traits and what trends do the progressions of these prices exhibit during the investigated timeframe?

Hypothesis:

Rising trends of all domestic housing market prices are expected, regardless of the geographical division or segments of the investigated timeframe.

Research limitations

- This research is based on online price listings made available by the seller/lessor himself (or parties representing him/her):
  - instead of being average prices, results are averages of what sellers/lessors perceive as appropriate prices for their properties;
  - not an ideal economic measure, however a rather accurate marker of where unobserved prices land at the time of transaction.
- Data proportionality with respect to the geographic locations from where it originates:
  - (in this case) the researcher does not get to pick the sample; the latter is determined by the amount of user-

Methods

The stages of gathering, organizing and transforming the raw microdata from the Internet are detailed below:

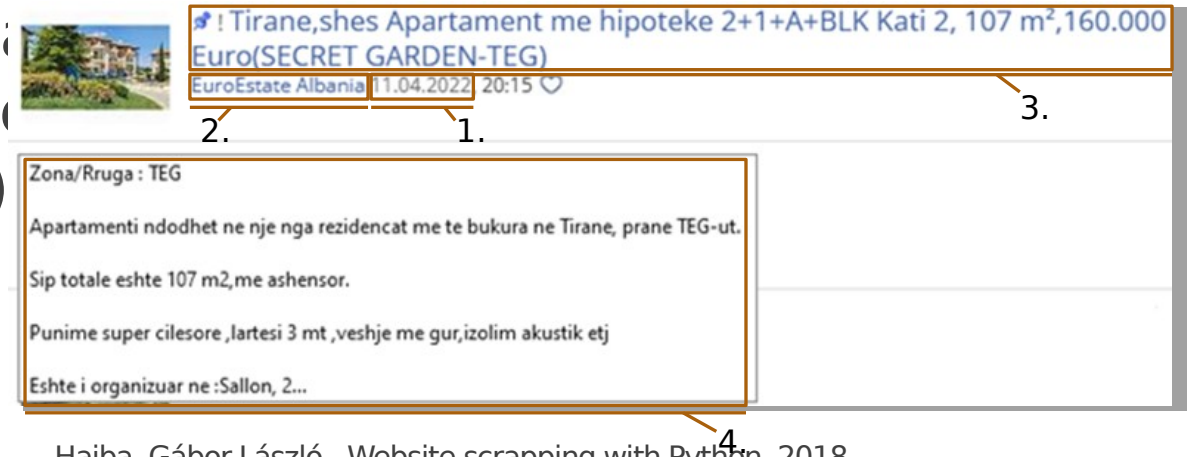
1. Identify the online platform where data resides and study its HTML structure, so as to obtain the snippets of information needed to be gathered for each listing (observation).
2. Use the Python environment (and its module Scapy) to code how the necessary information will be spotted in each web listing, and to collect all the data into a data interchange format file (JSON).
3. Access the available data through MS Excel, and use its module PowerQuery to bring it to tabular form, with defined variables, having neither duplicate values, missing values nor outliers.
4. Use PowerPivot to calculate price by sq.m. in a unified currency (ALL), as well as aggregate and average the data by month and geographic division, thereby obtaining 6 series from the two available samples of raw data.
5. In order to clearly distinguish price differences between the capital, the coast, and inland areas, use SAS' state-space models UCM procedure, and its Kalman algorithm, to conduct econometric regressions for estimating the series' trends.

Methods (cont'd)

The Python-executed data collecting methods used in this study are described in Hajba G. L. (2018)[a] and Kouzis-Loukas D. (2016).[b] The procedures consists of:

- identifying the website(s) containing the data (in our case: "njoftime.com");
- installing, configuring and using the Scrapy data collecting module within the Python environment;
- through Scrapy and Python tools, our case consisted of 4 steps of information collection, per observation, to be obtained:

1. Date (day, mnth., yr. when listing is added)
2. Listing agent (private individual / agency)
3. Header (property's most


! Tirane,shes Apartament me hipoteke 2+1+A+BLK Kati 2, 107 m²,160.000 Euro(SECRET GARDEN-TEG)
EuroEstate Albania
11.04.2022
20:15
1.
2.
3.
Zona/Rruga : TEG
Apartamenti ndodhet ne nje nga rezidencat me te bukura ne Tirane, prane TEG-ut.
Sip totale eshte 107 m2,me ashensor.
Punime super cilesore ,lartesi 3 mt ,veshje me gur,izolim akustik etj
Eshte i organizuar ne :Sallon, 2...
4.

(a. Hajba, Gábor László - Website scrapping with Python, 2018.
(b. Kouzis-Loukas, Dimitrios - Learning Scrapy: Learn the art of efficient web scraping and crawling with Python, 2016.

Methods (cont'd)

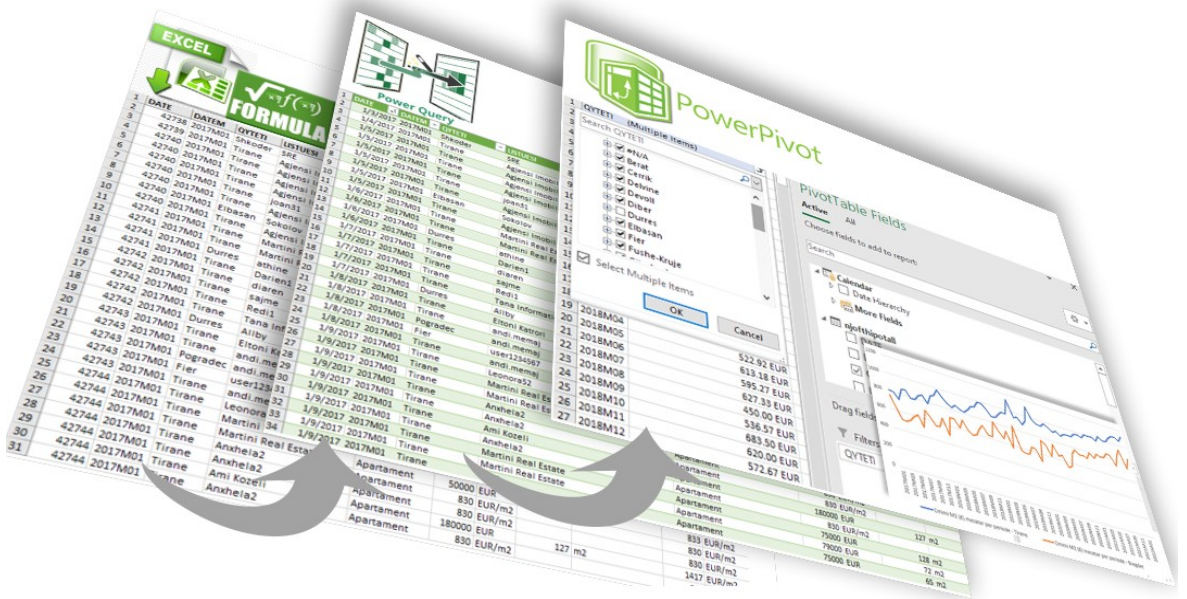
Python code example for Scrapy:

- The URL of the webpage to be rendered is specified at the beginning.
- Items such as “date”, “lister”, “header”, etc., search the price listings for the snippets of information to be collected.
- The data collection is

```
import scrapy


class shtepi_try2Spider(scrapy.Spider):
    name = 'shtepi_try2'
    allowed_domains = ['njoftime.com']
    start_urls = ['https://www.njoftime.com/forumdisplay.php?4-Shtepi-ne-shitje/page{}
    download_delay = 2

    def parse(self, response):
        for box in response.css('div.inner'):
            yield {
                'date': box.css('dd::text').get().replace(', ',''),
                'lister': box.css('dd div a strong::text').get(),
                'header': box.css('.title::text').get(),
            }

    def parse(self, response):
        for link in response.css('h3.threadtitle a::attr(href)'):
            yield response.follow(link.get(), callback=self.parse_info)

    def parse_info(self, response):

    def start_requests(self):
        for i in range (1, 1305):
            yield scrapy.Request(self.start_urls[0].format(str(i)), callback=self.parse)
```

Methods (cont'd)

- Python tools and code collect/save the raw microdata into a .JSON file.
- Excel formulas are afterwards applied to clean, organize and tabulate them.
- PowerQuery brings the data together into a single data framework/model.
- The data model is finally processed through PowerPivot and the DAX language into the needed results (obtained filtered sets,



Details on data

Source of microdata: web platform "njoftime.com".

Referring methodology for geographic division of data: Bank of Albania's "Real Estate Market" semiannual questionnaire.
Data were divided into three groups:

1. Capital
   Tirana Municipality (residences within a 5 km radius from downtown).
2. Coast
   Cities, towns and rural settlements on country's coasts: Durrës, Himarë, Kavajë, Sarandë, Shëngjin, Vlorë.
3. Inland
   Cities, towns and rural settlements located in country's inland: Berat, Cërrik, Delvinë, Devoll, Dibër, Elbasan, Fier, F-Krujë, Gjirokastër, Gramsh, Kamëz, Korçë, Krujë, Kuçovë, Kukës, Laç, Lezhë, Librazhd, Lushnjë, M. e Madhe, Mirditë, Pogradec, Përmet, Shkodër, Tepelenë.

Data comprised of two samples:

		1. Houses for sale	2. Houses for rent
Observations:		31,240	34,053
Pct. of obs. for:	Capital:	84.4%	95.2%
	Coast:	8.7%	2.2%
	Inland:	6.9%	2.6%
Sampling frequency:		monthly	monthly
Timespan:		Jan. 2017 - Jul. 2022	Jan. 2017 - Dec. 2022
Data points (months):		67	72
Avg. observations per month:		466	473

Details on data (cont'd)

Samples are deemed representative because:

- they meet the condition of random selection, and;
- for each period (month), are sufficiently numerous to assess the respective populations' distributions and other elements of central tendency.

Samples were refined to solely include:

- apartments of residential buildings;
- single and multi storey residential houses and villas.

Omitted observations consisted of:

- premises used to conduct activities unrelated to residential accommodation, (such as business, social, industrial and agricultural activities);
- residential properties sold in conjunction with sizable areas of land or other natural assets;
- residences for rent where weekly/daily rental payments were required

Methods (cont'd)

The goal of estimating the series' trends, that enable clear distinctions of prices among the market's geographical divisions, brings about the econometric context of this research.

SAS, its state-space models module and its UCM procedure employing the Kalman algorithm, is used to calculate the trends of monthly averages, estimated from the observations.

The specification of the unobserved components model for each of our series, required for the specific needs of our analysis, takes the form:

where: trend component;
seasonality;
cycle component;
error, NIID

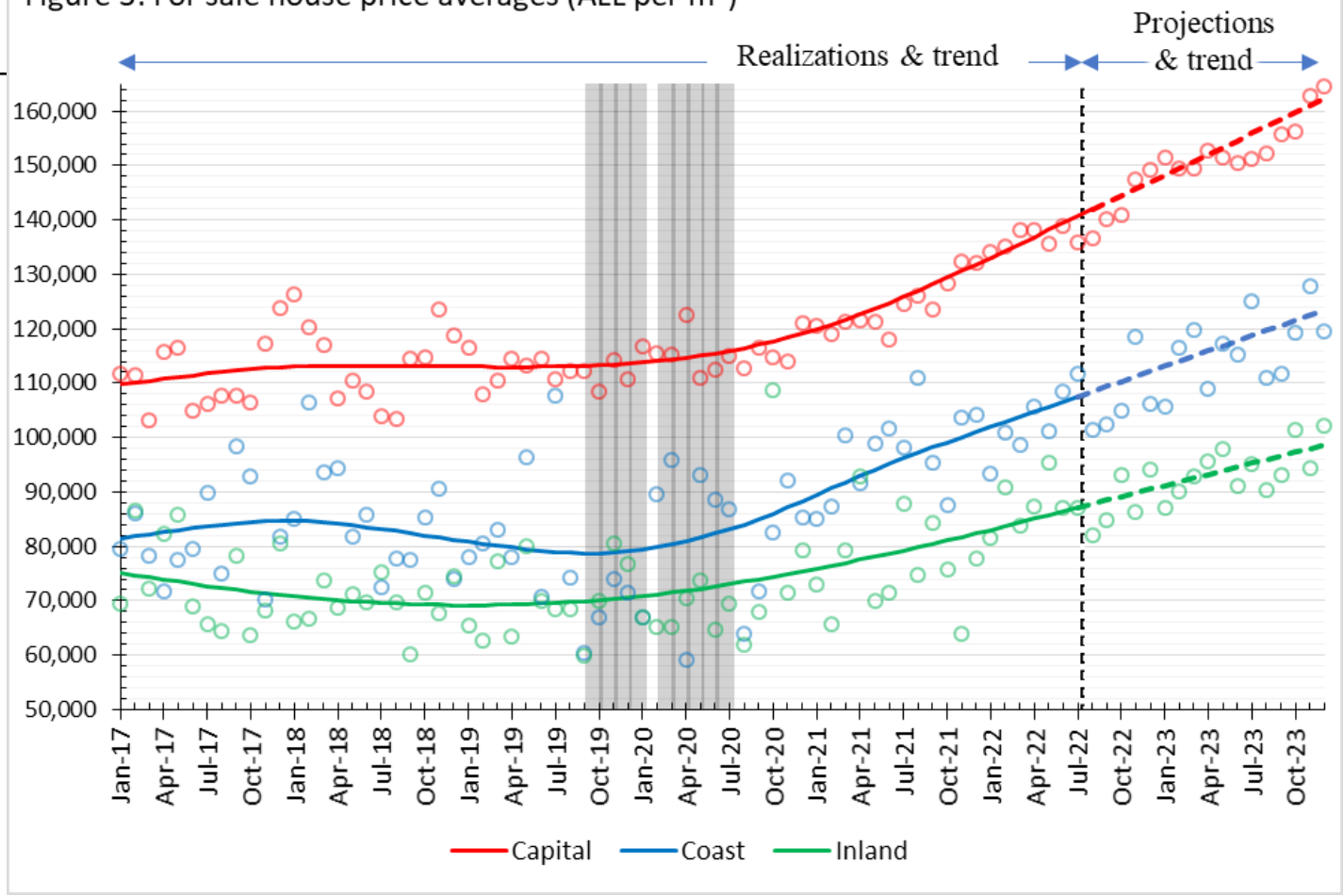
Worth noting that trend is composed of two sub-components of its own, level and slope , where:
and: NIID;
, and: NIID.

Results

Trends of per-sq.m. averages – houses for sale prices in ALL, for capital, coast, and inland areas.

- Considerable gap between capital and other areas; houses all over the country typically sell for ≈ ⅔-s of the prices level in the capital.
- Visible gap between coast and inland areas; houses located inland

Figure 3: For sale house price averages (ALL per m^2)



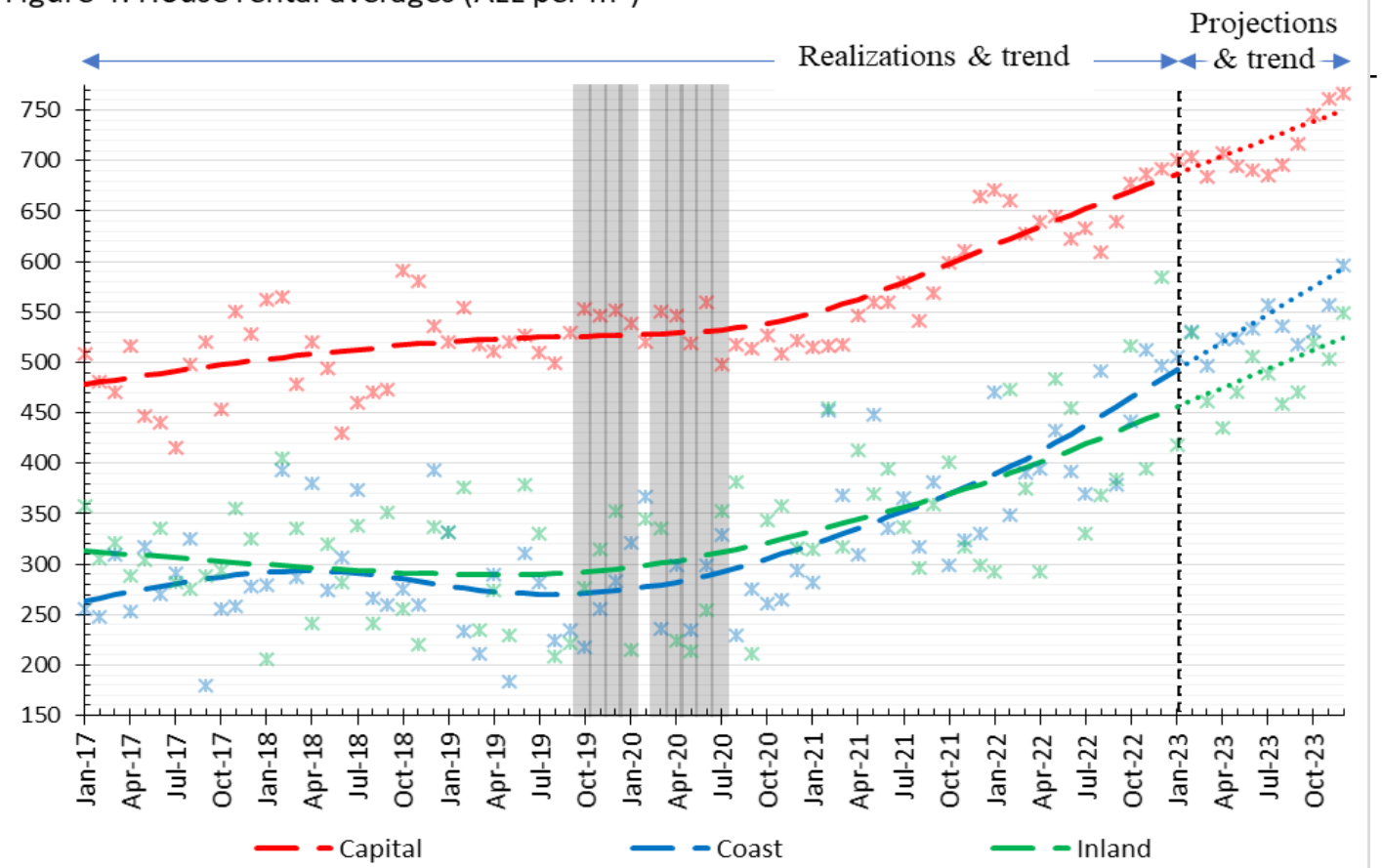
- Declining trend for prices in inland areas, 2017 - mid 2020.
- Accelerating upward trend for all prices, mid 2020 onward.

Results (cont'd)

Trends of per-sq.m. averages – rentals in ALL, for capital, coast, and inland areas.

- Considerable gap between capital and other areas; rentals all over the country typically fluctuate at ≈ ⅗-s of the rentals' level in the capital.
- Practically no gap between rentals on the coast, and rentals in inland areas.
- Declining trend for rentals in inland areas, 2017 - mid 2020.
- Accelerating upward trend for all rentals, mid 2020 onward.

Figure 4: House rental averages (ALL per m^2)



Results (cont'd)

Splitting the investigated period into two segments, [1] house prices' and [2] rentals' average YoY growth clips:

- during first segment (beginning of 2018 - mid 2020) were, respectively:
  - In the capital: 1.2% and 4.2%;
  - On the coast: -0.7% and 3.9%;
  - In inland areas: -1.8% and -1.8%.
- during the second segment (mid 2020 - end of 2022) were, respectively:
  - In the capital: 8.8% and 9.1%;
  - On the coast: 15.1% and 22.8%;
  - In inland areas: 11.9% and 22.2%.

Results (cont'd)

Between beginning of 2017 and mid 2020:

- House prices progression:
  1. In the capital: 3.1% rise
  2. On the coast: 5.9% rise
  3. In inland areas: 1.1% decline
- Rentals progression:
  1. In the capital: 9.9% rise
  2. On the coast: 9.7% rise
  3. In inland areas: 0.2% decline

Between mid 2020 and end of 2022:

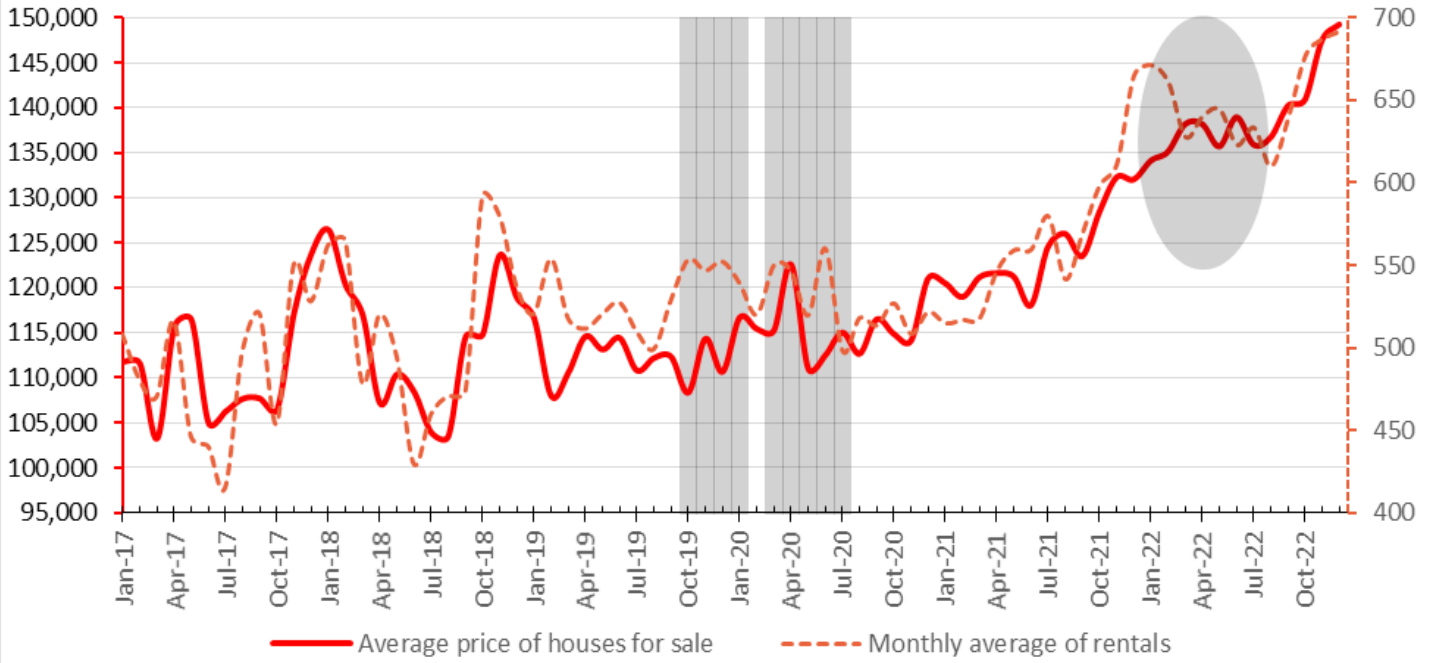
- House prices progression:
  1. In the capital: 28.3% rise
  2. On the coast: 29.4% rise
  3. In inland areas: 23.1% rise
- Rentals progression:
  1. In the capital: 29% rise
  2. On the coast: 68.6% rise
  3. In inland areas: 46.3% rise

Per sq. m. price averages s (in ALL), by December of year: (UCM forecast)	2022	2023
1. In the capital:	147,033	162,233
2. On the coast:	112,208	123.430
3. In inland areas:	90,522	98,685

Results (cont'd)

- House prices and rentals in the capital display notable covariation in time.
- There seem to be short time instances when these two series seem to move in opposite directions, particularly during times of economic distress (2019T4, 2020T2, 2022T1).
- In general their

Figure 5: House market price dynamics in the capital (ALL per m^2)



- During the period when armed conflict in Europe began, followed by international sanctions and subsequent rise of energy prices, the housing market reacted through a moderate increase of house prices and a notable decline of rentals in the capital.

Final considerations

In general terms, the residential housing main indicators display continuous growth in all geographical locations of Albania;

- This growth is particularly evident in coastal areas and the capital, but it is present in inland areas too.
- If we divide the investigated period into two segments (2017 – mid 2020 and mid 2020 - 2022), we note that:
  - for the first segment, house prices and rentals in inland areas show moderate declines;
  - for the second segment, results show accelerated double-digit growth trends on the coast and in inland areas.
  - especially prices on the coast appear to be on a quite fast catch-up path with the level of prices in the capital.

Final considerations (cont'd)

- Because of the prices' falling trends in inland areas, during the first segment (2017 – mid 2020), we conclude that the results do not support our research hypothesis.
- Meanwhile, growth trends in the capital, even though still high, remained in single digits, which means that housing prices' gap between the capital and the rest of the country is getting narrower.
- Even though it is still early to conclude, these seem to be the first indications of a housing prices' trend plateau in the capital, which should become more apparent in the months/quarters to come.

Future research

The big data approach to microeconomic research in the Bank of Albania currently is in its embryonic stage.

- User-generated content (UGC) constitutes only a small portion of the immense volume of Internet-residing information known as “big data”, but it leads to tremendous opportunities in the field of research.
- An appealing research frontier, where UGC microdata related to the housing market can be used, is hedonic pricing. It relies on a given product’s characteristics to account for the variation in its declared price, and UGC can be an ample source of information on product characteristics.

Thank You

… for your attention!

Questions, comments, observations appreciated.